Bi-national bi-modal bi-lingual corpora of child language

Ronice Müller de QUADROS, Diane LILLO-MARTIN, Deborah CHEN PICHLER
Universidade Federal de Santa Catarina; University of Connecticut; Gallaudet University
ronice.quadros@ufsc.br, lillo.martin@uconn.edu, deborah.pichlerr@gallaudet.edu

Abstract

This paper discusses projects involving the building of corpora of sign language acquisition data. We developed a methodology to collect, to transcribe and to store data from different contexts of acquisition. The corpora include deaf children, from deaf parents; deaf children, from hearing parents; hearing children, from deaf parents (Codas) and deaf children with cochlear implants. There are two sign languages involved: Brazilian Sign Language and American Sign Language and two spoken languages, in the bilingual bimodal cases, that are, Brazilian Portuguese and American English. The complexity of building these corpora includes development of patterns of transcription and the organization of the same metadata system. In this process, we are developing manuals, database and software to make the data available and comparable across the languages. One example of software that we present in this paper concerns Sign ID, that is, it is software to indicate identities for each sign that is part of the database. The Sign ID software helps us make the annotations more consistent across transcribers. This kind of work is making it possible to compare data from these languages.

Keywords: sign language; corpora; and language acquisition.

1. Introduction

In order to address numerous linguistic research questions, we have been building several corpora of sign language acquisition data. Until recently, our focus had been on sign language only with deaf children, from deaf parents, acquiring sign language as native language. In this case, we built corpora of longitudinal data collected over a long period of time: these corpora included spontaneous data, with interaction of the child from 1-4 years old and an adult (usually the Deaf mother or a Deaf experimenter),. On the Brazilian side, there is also data from deaf children with hearing parents. In this context, a Deaf experimenter interacts with the child in sessions alternating with the hearing mother. All the analyses done so far indicate that in the specific context of deaf children with deaf parents, the sign language acquisition is parallel to spoken language acquisition (see Lillo-Martin, 1999 and Newport & Meier, 1985 for reviews of some of this). However, there are also findings showing that certain aspects of language acquisition in this context show modality effects (e.g. Meier & Newport, 1990; Marentette & Mayberry, 2000; Meier, 2006). On the other hand, in the context in which the deaf child has limited contact with sign language, there is a lot of variability in the language development reported by different researchers, but it seems that even in these contexts in which *input* is not conventional, because the child has parents learning sign language and restricted or no access to sign language, the child develops his/her signing skills better than his/her parents, showing that the child is able to make better use of the mental language system (e.g. Singleton & Newport, 2004; Goldin-Meadow, 2003; Goldin-Meadow & Mylander, 1984, 1990, 1998; Quadros & Cruz, 2011).

Now we are expanding our work to include bimodal bilingual children acquiring both a sign language and a spoken language, building comparable corpora across two sign/spoken language pairs: Brazilian Sign Language and Brazilian Portuguese on the one hand, and American Sign Language and American English on the other. We are again collecting longitudinal data with babies from 1 to 4 years old, and adding experimental data with children from 4 to 7 years old.

We use different sets of researchers (deaf and hearing) to emphasize appropriate target language use, assuming the child's interlocutor sensitivity (Petitto *et al*., 2001), but we also recognize that code-blending is simply a part of the language system being acquired.

We reorganized the form of the database used with the longitudinal data and we built a new database for the experimental studies. The experimental studies include a set of 24 tests, evaluating different language aspects, such as, morphology, phonology, syntax, discourse and pragmatics. The goal of the tests is to provide a comprehensive profile of each bilingual child's developing competency in Libras (Brazilian Sign Language) and Brazilian Portuguese, or ASL (American Sign Language) and American English.

The data in sign and in speech adds considerable complexity to the already challenging prospect of corpus building. In this presentation, we explore some of the issues we have faced already and those we expect to face, in the context of our linguistic goals.

Recent research on childhood bilingualism has indicated that although children have two separate developing grammatical systems from very early on, there are instances of cross-linguistic influence, where grammatical structures from one language seem to exert a temporary influence on the child's grammar of the other language (e.g. Hulk & Müller, 2000). An important question is to identify the loci of such influences based on linguistic criteria. In order for us to address such issues, we are developing corpora from individual children acquiring both a sign language and a spoken language. Many of the same data collection issues arise as those for projects investigating only sign language (see Baker & Woll, 2005 for some best practices in this domain). However, in our current project, it turns out that there are specific things for which additional practices are needed; for instance, we frequently observe code-blended

Heliana Mello, Massimo Pettorino, Tommaso Raso (edited by), *Proceedings of the VIIth GSCP International Conference : Speech and Corpora*
ISBN 978-88-6655-351-9 (online) © 2012 Firenze University Press.

language (the use of signs and speech produced simultaneously) as well as unimodal productions (Bogaerde & Baker, 2005, 2009; Emmorey *et al*., 2008). Language- or modality-specific properties as well as universals are found to be very interesting in these contexts. In this paper, we will present the organization of the sign language acquisition corpora developed on both sides of the project: Brazil and the United States of America.

2. Metadata

The metadata of the children is organized through documents that are shared with researchers involved in the different steps of the investigation: data collection involving filming, transcribers, people that organize the data for specific purposes and people that analyse the findings. The main topics of the documents are the following:

LONGITUDINAL

- Protocol of the child (nickname of the child, for example, EDU)
- Number of the section (from 000 up to the number of the sections collected, for example, EDU001, EDU002, EDU003)
- Date of the filming
- Age of the child (years;months.days)
- Target language
- Duration of the session
- Adults involved in the session
- Other participants involved in the session
- Comments
- Transcribers
- Checker/reviser of the transcription
- Organizer of the data for each purpose (for example, for WH analysis, for Modality analysis, etc.)

EXPERIMENTAL

- Name of the test
- Nickname of the child
- Condition (Coda, Deaf, CI, Coda adult)
- Date
- Age
- Language
- Duration
- Comments
- Transcriber
- Reviser

The whole database is organized in a computer server. See Figure 1 for an illustrative sample of this organization. There are two main folders: the original archive ("*acervo*") and the production. The first one has the original videos. The second one has the compressed videos for manipulation by the people that access the videos, as well as transcription and analysis files.

The production folder includes the experimental data and longitudinal data in separate sections. First we discuss the longitudinal data. The basic organization is to list the children in separate folders. Each child's folder will include the folders for each session containing the video and the transcript files (the basic one and the ones with the specific organization for specific purposes). The transcription is done using ELAN software producing eaf files with separate tiers of annotation capturing different types of information (see also below).

For the experimental studies, the basic organization is to have the folders with the places and years in which the fairs happened. Within each place, the folders are separated by test. These folders are further divided into two sets of data by child: one for those whose data is without restriction *("sem restrição")*, and another for restricted data ("*com restrição*"). The restrictions are related to the kind of access people have to the videos. Some of the parents do not want students to have access to the videos of their child or for the researchers to use frames of the videos in conferences, for example. Within these two folders based on restriction, the children, then, are listed with the video and the eaf or the form of the test scanned with the results, depending on each test.

In the case of the experimental studies, the database is organized as well as using FileMakerPro (Figure 2 in Appendix). This database includes all four languages. Then, it facilitates the comparison among the experimental results over the four languages.

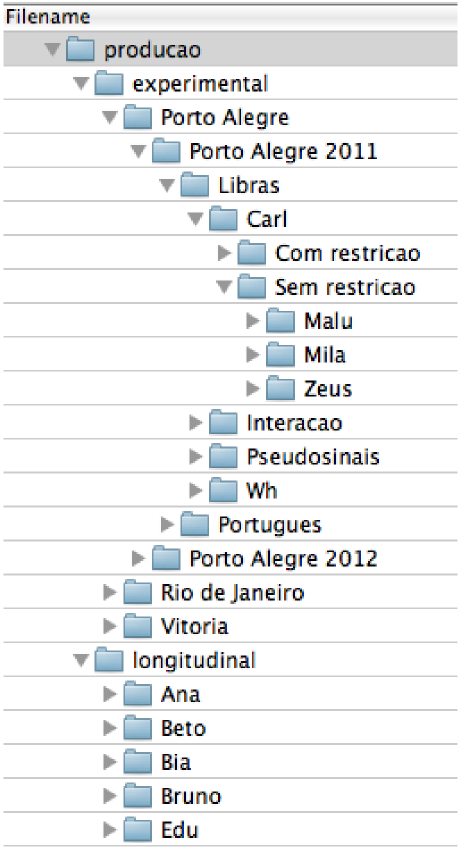


Figure 1: Example of the organization of the database

3. Designing annotation patterns

Following video collection, we invest considerable energy in the production of transcripts, to be used in conjunction with the videos for linguistic analyses. Following our earlier sign-only research, we use ELAN for time-locked videos with transcription (http://www.latmpi.eu/tools/elan/).

For bilingual research, we designed a different template so that both languages are parent tiers, to optimize the study of (sequential or simultaneous) bimodal productions. See Chen Pichler *et al*. (2010) for a detailed description of our ELAN tier structure and transcription conventions (cf. Figure 3 and Figure 4, in Appendix).

The general principles that guide the annotation of the data are to create a machine-readable record of language samples, not necessarily sufficient for the reader to reproduce in exactly the same way, but so that the records can be searched to find all occurrences of phenomena of interest (in the way described by Johnston, 2001, Johnston & Schembri, 2007; Miller, 2001; Pizzuto & Pietrandrea, 2001). In addition to having a basic annotation of the utterance in each language, we use multiple annotation parses focusing on different phenomena. This documentation of the data is the foundation for our analysis decisions.

Where it is possible, we follow the CHILDES conventions established for child language data (MacWhinney, 2000) in transcribing both speech and sign (though we do not use BTS) http://childes.psy.cmu.edu/manuals/chat.pdf. When the CHILDES conventions conflict with our sign-specific goals, we create new conventions to be followed for transcribing both sign and speech. It is important to keep the sign and speech transcriptions comparable.

4. Sign IDs

Finally, we see a number of important implications and extensions of the system we are developing. For example, we are creating a specific identification for each sign to be used in our transcripts (in the same spirit of Johnson, in preparation, for Australian Sign Language), what we call “Sign ID”. Because there is no commonly accepted writing system for sign languages, sign researchers generally rely on a system of glossing; however, traditional transcription does not assign a consistent gloss for each sign, but different glosses depending on context and other aspects of the signed utterance. This means that it is very difficult for researchers to identify the locations of interest in a transcript using a search function to discover all occurrences of a particular sign. Analysis must proceed at a much slower pace of hand searching transcripts one utterance at a time. In order to facilitate and expand the analysis of data collected in the parent project, we developed a sign ID lexicon containing the vocabulary items used most frequently by the children we are studying. Sign IDs are word labels chosen to represent each sign root systematically, so that every use of the sign has the same label, despite contextual or morphological differences which affect how the sign is interpreted. By using sign IDs in our transcripts, we are able to conduct our analyses more efficiently, using a wider range of data. The sign ID lexicon addresses the problem of transcript searchability and greatly facilitates the analysis of data collected for sign language corpora. This helps to standardize annotations and it can be more freely accessed by other researchers.

On the Brazilian side, we have been developing the sign IDs database by feeding it with the signs over which transcribers had doubts regarding transcription. We have periodic meetings to discuss these signs, then we christen each and add it to the ID list (www.idsinais.libras.ufsc.br) (see Figure 5 in Appendix for the Sign ID screen). The search system has filters based on sign language parameters (132 handshapes divided in 13 groups and 8 locations). An example with a group of handshapes chosen as a parameter to search for a specific sign is given in Figure 6 and the results of this search are shown in Figure 7, in Appendix.

The sign ID specifications include identification of the sign, Portuguese translation, English translation, written sign, handshape groups, handshapes, location and sign video. The searching may be done through handshapes, locations, handshape groups, location groups, the sign ID or the first letter of the sign ID.

On the American side, the development of an ID gloss database has taken into consideration the needs of different research groups across the country, each of which uses a different system for writing signs. The database was set up so that different local groups can enter their own information about each sign, and each group can also view the information entered by the others. This approach will facilitate the comparison of transcriptions used across different groups, and may eventually lead to greater convergence in the glossing systems used.

5. Conclusion

One of our major goals has been cross-site comparability, that is, establishing the same criteria, approach to data collection, ELAN template, and general transcription principles to be used across our three universities. The metadata and data are shared through the use of a common server, as well as online services including Google docs and Dropbox. The analyses of the results are being conducted through regular meetings and we are on the right track to answer our research questions (e.g., Lillo-Martin *et al*., 2010; Chen Pichler *et al*., 2010; Quadros *et al*., in press).

We have not yet resolved the following linguistic issues, but we hope that our project will contribute to their discussion in the field as a whole. Does bimodal bilingualism lead to cross-language influence different from that found in mono-modal bilingualism (e.g., due to code-blending, or use of non-manuals)? When bimodal bilinguals code-blend, are they choosing grammatical structures which are permitted in both languages for maximum accommodation? What kinds of syntactic representations can account for code-blends? These are the types of research questions our project can address through the use of the corpora we are now building.

Our template and corpus-building decisions can be applicable to the development of adult only bimodal bilingual corpora. In addition, many similar issues arise in the study of co-speech gesture, and researchers in this area may take advantage of aspects of our procedures. And, we hope that our collaboration across continents may contribute to and promote cross-linguistic research on sign languages as well.

6. Acknowledgements

This research is supported by the U.S. National Institutes of Health – NIDCD grant #DC00183 and NIDCD grant #DC009263; by a Gallaudet University Priority Grant; and by the Brazilian National Council for Research, CNPq Grant #CNPQ # 304102/2010-5 and # 471478/2010-5.

We sincerely thank the Deaf consultants, research assistants, children, and their families who work with us in our research.

7. References

Baker, A., Woll, B. (Eds.) (2009). Sign language acquisition. Amsterdam: John Benjamins.

Bogaerde, B. van den, Baker, A.E. (2005). Code-mixing in mother-child interaction in deaf families. In *Sign language & linguistics*, 8(1-2), pp. 151--174.

Bogaerde, B. van den, Baker, A.E. (2009). Bimodal language acquisition in Kodas (kids of deaf adults). In M. Bishop, S.L. Hicks (Eds.), *Hearing, mother-father Deaf: Hearing people in Deaf families*, Washington, DC: Gallaudet University Press, pp. 99--131.

Chen Pichler, D., Hochgesang, J., Lillo-Martin, D. and Quadros, R. M. (2010). Conventions for sign and speech transcription of child bimodal bilingual corpora in ELAN. In *Language, Interaction and Acquisition, 1*, pp. 11--40.

Chen Pichler, D., Quadros, R.M. and Lillo-Martin, D. (2010). Effects of Bimodal Production on Multi-Cyclicity in Early ASL and LSB. In J. Chandlee, K. Franich, K. Iserman, and L. Keil (Eds.), *A Supplement to the Proceedings of the 34th Boston University Conference on Language Development*. Available at: <www.bu.edu/linguistics/BUCLD/supp34.html>.

Emmorey, K., Borinstein, H.B., Thompson, R. & Golan, T.H. (2008). Bimodal bilingualism. In *Bilingualism: Language and cognition*. 11(1), pp. 43--61.

Goldin-Meadow, S. (2003). *The resilience of language: what gesture creation in deaf children can tell us about how all children learn language*. New York: Psychology Press.

Goldin-Meadow, S., Mylander, C. (1984). Gestural communication in deaf children: The effects and noneffects of parental input on early language development. Monographs of the Society for Research in Child Development, 49 (3–4, Serial No. 207).

Goldin-Meadow, S., Mylander, C. (1990). Beyond the input given: The childs role in the acquisition of language. In *Language*, 66, pp. 323--355.

Goldin-Meadow, S., Mylander, C. (1998). Spontaneous sign systems created by deaf children in twocultures. In *Nature*, 391, pp. 279--281.

Hulk, A., Müller, N. (2000) Bilingual first language acquisition at the interface between syntax and pragmatics. In *Bilingualism: Language and Cognition* 3 (3), 2000, Cambridge University Press, pp. 227--244.

Johnston, T. (in preparation). *From archive to corpus: Transcription and annotation in the creation of signed language corpora, manuscript*. Department of Linguistics, Macquarie University, Australia.

Lillo-Martin, D. (1999). Modality effects and modularity in language acquisition: The acquisition of American Sign Language. In T. Bhatia & W.C. Ritchie (Eds.), *Handbook of Language Acquisition*, San Diego: Academic Press, pp. 531--567.

Lillo-Martin, D., Quadros, R.M., Koulidobrova, H. and Chen Pichler, D. (2010). Bimodal Bilingual Cross-Language Influence In Unexpected Domains. In J. Costa, A. Castro, M. Lobo and F. Pratas (Eds.), *Language Acquisition and Development: Proceedings of GALA 2009*, Newcastle upon Tyne: Cambridge Scholars Press, pp. 264--275.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Marentette, P., Mayberry, R. (2000). Principles for an emerging phonological system: A case study of acquisition of American Sign Language. In C.D. Chamberlain, J.P. Morford and R. Mayberry (Eds.), *Language Acquisition by Eye*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 51--69.

Meier, R. (2006). The form of early signs: Explaining signing children's articulatory development. In B. Schick, M. Marschark and P.E. Spencer (Eds.), *Advances in Sign Language Development by Deaf Children*, New York: Oxford University Press, pp. 202--230.

Meier, R.P., Newport, E.L. (1990). Out of the hands of babes: On a possible sign advantage in language acquisition. In *Language*, 66, pp. 1--23.

Newport, E.L., Meier, R.P. (1985). The acquisition of American Sign Language. In D.I. Slobin (Ed.), *The Cross-Linguistic Study of Language Acquisition, Volume 1*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 881--938.

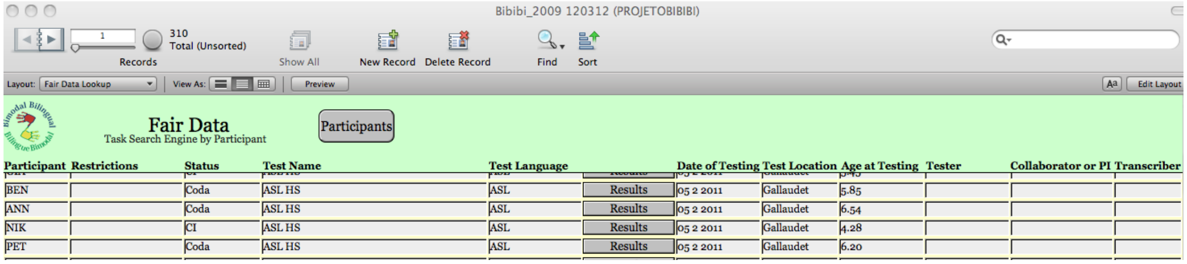
Petitto, L.A., Katerelos, M., Levi, B., Gauna, K., Tetrault, K. and Ferraro, V. (2001). Bilingual signed and spoken language acquisition from birth: Implications for the mechanisms underlying early bilingual language acquisition. In *Journal of child language*. 28(2), pp. 453--496.

Quadros, R.M., Lillo-Martin, D. and Chen Pichler, D. (in press). Early effects of bilingualism on WH-question structures: Insight from sign-speech bilingualism. In *Proceedings of GALA 2011*. Newcastle upon Tyne:

Cambridge Scholars Press.

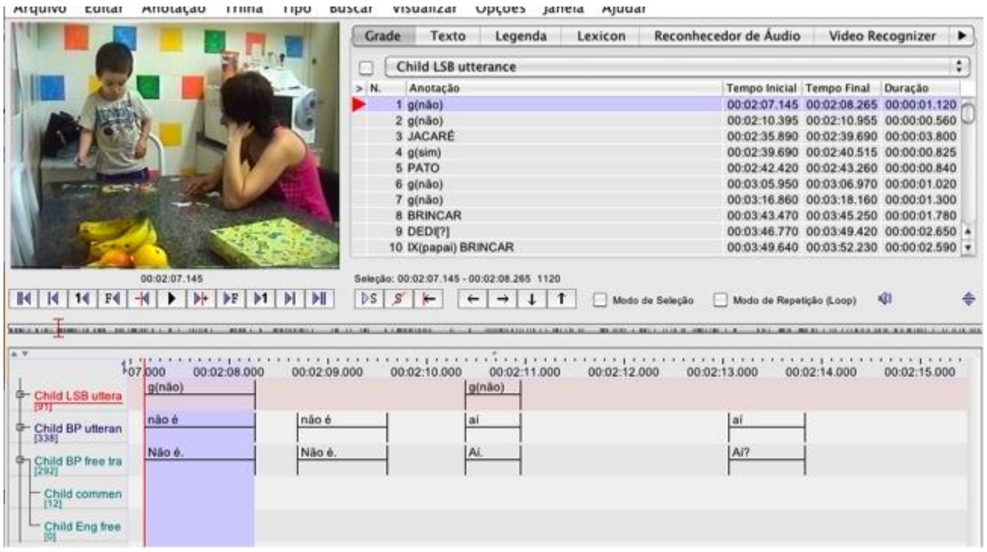
Singleton, J.L., Newport, E., (2004) E. When learners surpass their models: The acquisition of American Sign Language from inconsistent input. In *Cognitive Psychology* 49. pp. 370--407.

8. Appendix



Participant	Restrictions	Status	Test Name	Test Language		Date of Testing	Test Location	Age at Testing	Tester	Collaborator or PI	Transcriber
BEN		Coda	ASL HS	ASL	Results	05 2 2011	Gallaudet	5.85			
ANN		Coda	ASL HS	ASL	Results	05 2 2011	Gallaudet	6.54			
NIK		CI	ASL HS	ASL	Results	05 2 2011	Gallaudet	4.28			
PET		Coda	ASL HS	ASL	Results	05 2 2011	Gallaudet	6.20			

Figure 2: FileMakerPro

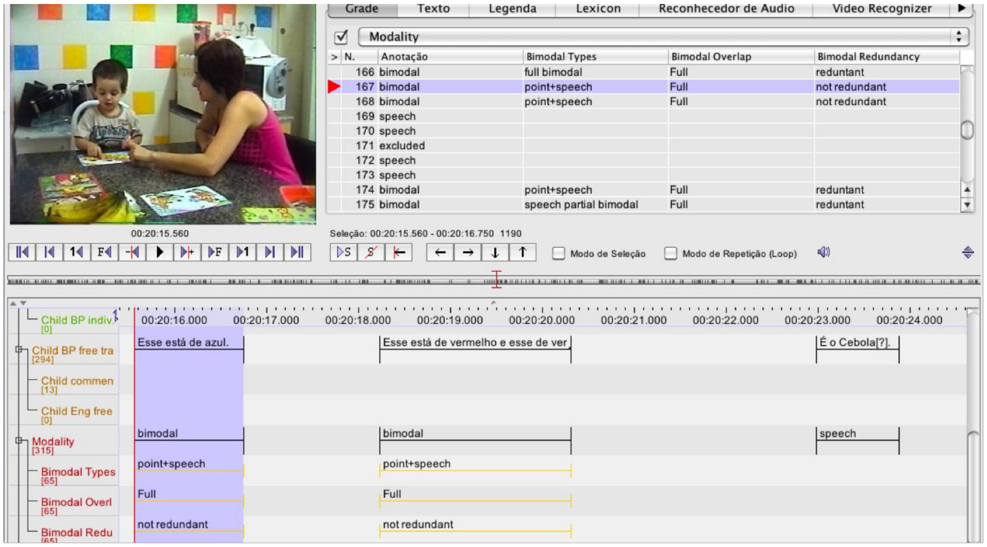


☐ Child LSB utterance

N.	Anotação	Tempo Inicial	Tempo Final	Duração
1	g(não)	00:02:07.145	00:02:08.265	00:00:01.120
2	g(não)	00:02:10.395	00:02:10.955	00:00:00.560
3	JACARÉ	00:02:35.890	00:02:39.690	00:00:03.800
4	g(sim)	00:02:39.690	00:02:40.515	00:00:00.825
5	PATO	00:02:42.420	00:02:43.260	00:00:00.840
6	g(não)	00:03:05.950	00:03:06.970	00:00:01.020
7	g(não)	00:03:16.860	00:03:18.160	00:00:01.300
8	BRINCAR	00:03:43.470	00:03:45.250	00:00:01.780
9	DEDI[?]	00:03:46.770	00:03:49.420	00:00:02.650
10	IX(papai) BRINCAR	00:03:49.640	00:03:52.230	00:00:02.590

Seleção: 00:02:07.145 - 00:02:08.265 1120

Figure 3: ELAN in the context of Bibibi Project with the basic tiers for the child



☑ Modality

N.	Anotação	Bimodal Types	Bimodal Overlap	Bimodal Redundancy
166	bimodal	full bimodal	Full	reduntant
167	bimodal	point+speech	Full	not redundant
168	bimodal	point+speech	Full	not redundant
169	speech			
170	speech			
171	excluded			
172	speech			
173	speech			
174	bimodal	point+speech	Full	reduntant
175	bimodal	speech partial bimodal	Full	reduntant

Seleção: 00:20:15.560 - 00:20:16.750 1190

Figure 4: ELAN in the context of Bibibi Project with the specific tiers for modality analysis

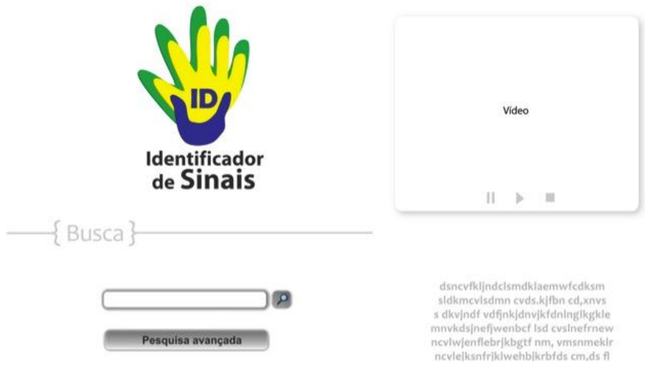

ID
Identificador de Sinais
{ Busca }
Vídeo
Pesquisa avançada
dsncvfkljndclsmdklaemwfcdksm
sldkmcvlsdmn cvds.kjfbn cd,xnvs
s dkvjndf vdfjnkjdnvjkfdnlnglkgkle
mnvkdsjnefjwenbcf lsd cvslnefrnew
ncvlwjenflebrjkbgtf nm, vmsnmeklr
ncvlejksnfrjklwehbjkrbfds cm,ds fl

Figure 5: ID screen for Libras

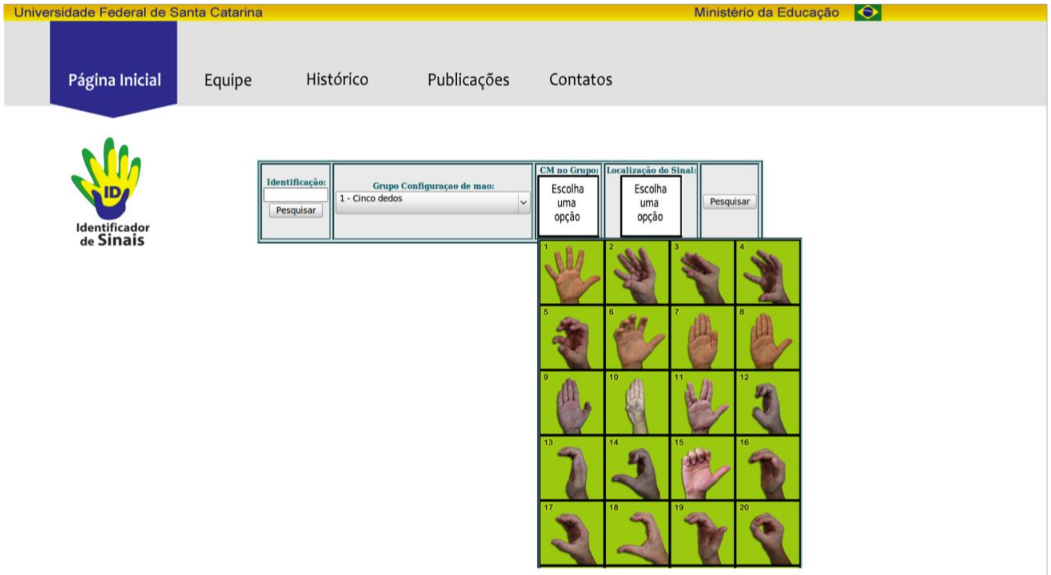


Figure 6: ID searching system: Handshape selection

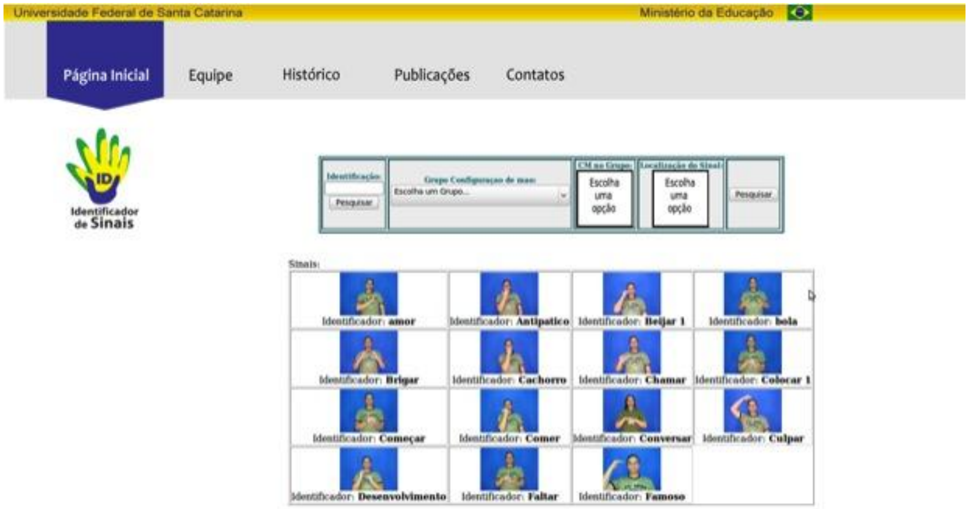


Figure 7: ID result of a search